INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

PROYECTO IA

WILSON ARBEY CHAVERRA MEJÍA
CC: 8064138
INGENIERÍA MECÁNICA

ALVARO JAVIER TALAGUA ROSARIO
CC: 1196969101
INGENIERÍA MECÁNICA

UNIVERSIDAD DE ANTIOQUIA
MEDELLÍN
2023

Home Credit Default Risk

Enlace de la competición en Kaggle:

https://www.kaggle.com/c/home-credit-default-risk/overview

Enlace del video:

https://drive.google.com/file/d/1QeBujWWbKY_-ifwsNrNa-C2bEPrh1-5g/view

El propósito del proyecto Home Credit Default Risk consiste en anticipar las posibilidades de que los individuos que solicitan créditos no cumplan con sus obligaciones de pago. Con el fin de alcanzar este objetivo, es de suma importancia llevar a cabo una adecuada preparación de los datos previos a su análisis y modelados. Este informe presenta un progreso en la manipulación de datos que puede resultar en una mejora sustancial en la precesión de las predicciones.

El proceso inicial en la manipulación de datos implica cargar los archivos de datos de entrenamiento y prueba. Esta acción se lleva a cabo mediante la utilización de la biblioteca Pandas de Python, aprovechando la función “read_csv”. A continuación, los conjuntos de datos se fusionan en un único DataFrame por medio de la función “concat”. Esta consolidación facilita una manipulación de datos más coherente y eficiente. Una vez que los datos se han fusionado, se procede a realizar algunas modificaciones fundamentales.

En este avance se centra en la creación de variables nuevas a partir de variables preexistente, específicamente, se generan dos variables adicionales “income_per_person” y “credit_term”. La primera variable se obtiene al dividir el ingreso total entre el número de miembros en la familia, mientras que la segunda variable se calcula al dividir el monto del préstamo entre el período en días. Esta transformación de datos aporta una mayor profundidad al análisis.

La incorporación de estas variables recién creadas puede significar una mejora sustancial en la precisión de las predicciones con respecto al incumplimiento de pago. En particular, la variable “income_per_person” se presenta como un indicador sólido de riesgo de incumplimiento, dado que los hogares con ingresos más bajos suelen más inconvenientes para cumplir con sus obligaciones financieras. Por otro lado, la variable “credit_term” puede ser la utilidad para identificar a los solicitantes que optan por plazos de pagos excesivamente cortos o largos, lo que podría indicar un mayor riesgo de incumplimiento, no obstante, es crucial subrayar que la creación de nuevas variables debe efectuarse con mucho cuidado y basándonos en lo que sabemos sobre el tema. A demás, debemos asegurarnos de que estas nuevas variables realmente aporten información valiosa al modelo que utilizamos para hacer predicciones.

Otro avance clave en el manejo de datos es lidiar con los valores que faltan, prácticamente en todos los conjuntos de datos, se encuentran valores que hacen faltan, lo que puede ser un problema para los modelos de predicción si no se abordarán adecuadamente. En este proyecto, se optó por usar una técnica de imputación basada en la media y la mediana para llenar los valores faltantes de los datos. Esta técnica de imputación basada en la media y la mediana es un método sencillo pero efectivo para lidiar con los datos faltantes. Básicamente, implica sustituir los valores restantes con el valor promedio o la mediana de los valores disponibles en la misma columna. Esto asegura que los datos estén completos y evita introducir sesgos o errores en el modelo de predicción.

Es fundamental recordar que la técnica de imputación basada en la media y la mediana no es apropiada para todos los conjuntos de datos. En algunas situaciones, puede ser necesario emplear métodos mas avanzaos, como la imputación múltiple.

En resumen, el proceso de manipulación de datos desempeña un papel esencial en la creación de modelos de predicción de riesgo crediticio, como el proyecto home Credit Default Risk. En este informe, se han presentado varios avances de manipulación de datos, que incluye la creación de nuevas variables, la imputación de valores restantes y la eliminación de valores que no son relevantes. Estos avances tienen como propósito mejorar la precisión de las predicciones y, en consecuencia, contribuir a la detección temprana de solicitantes de créditos con un alto riesgo de incumplimiento.

Es esencial recordar que la manipulación de datos requiere un enfoque cauteloso y respaldado por el conocimiento experto en el campo. Se deben aplicar las técnicas adecuadas y monitorear constantemente la calidad y efectividad de estas técnicas. A demás es fundamental llevar a cabo una validación rigurosa del modelo para asegurarse de que las predicciones sean precisas y confiables.

En conclusión, los avances en la manipulación de datos que se han presentado en este informe desempeñan un papel fundamental en el éxito del proyecto Home Credit Default Risk. Al mejorar la calidad de las predicciones, se anticipa que estos avances contribuirán a detectar tempranamente a los solicitantes de crédito con un alto riesgo, lo que, a su vez, puede ayudar a reducir el riesgo de impago de los préstamos y mejorar la rentabilidad de la compañía.